**IBM Business Attrition- Exploratory Data Analysis**

**Project Status Report**

**Abstract**

Employee attrition is when employees leave the company. Towards Data Science (2018) “attrition in human resources refers to the gradual loss of employees over time”. This type of analysis considers retirement and resignation and looks for trends to retain the best employees and understand if there are variables affecting the retention in the organization.

The goal with this project is to predict employee attrition. We have used different predictive and correlation functions to understand and predict employee attrition, as well as determine key variables affecting attrition.

**Introduction**

HR analytics (Human resource analytics) is the usage of analytics and analytical tools to support Human Resources processes in organizations; it allows to analyse, measure, control, and improve employees’ performances, as well as overall results and return of investment related to human capital. All companies could use HR analytics, but it is even more important for big companies because they need to analyse all variables to be able to design compensations and motivational programs to retain the best employees. But how do the companies know which are the best employees? The HR analytics can help measuring the results and performance of employees to analyse what type of employees they want to retain, who has potential, and how can let go.

In addition, HR analytics measures the results in terms of planning, execution and potential for each employee to use this data to support decisions relating with compensations, investment, development, as well as assigning employees to different groups and detect the most talented, or top players/employees.

**Project Overview**

We decided to analyze the database of IBM HR Analytics Employee Attrition & Performance dataset available in Kaggle, and we will analyze information that measure the connection between attrition and other variables, like age, job satisfaction, education, years in company, etc. We used different packages and commands in RStudio to measure the correlation and predict the attrition for IBM.

Our research questions are:

* How does years since last promotion impact employee’s attrition?
* Is the turnover higher for those employees travelling more?
* What is the relationship between education field or job role and attrition?
* How does Monthly income affect employee attrition?
* What are the most important predictors for the employee attrition? Can we find correlation between the predictors?
* What are the major differences between the employees who are likely to leave and who are likely to stay?
* Is there a model that help us to predict the future employee attrition?

The key ideas we will look to review with this project

* Correlation between employee performance and possible attrition
* Trends in employee attrition with increasing duration in the company
* Correlation between employee satisfaction and possible attrition

**Related Works**

We found different works using either the same data source or analyzing the same issue.

* Sharma, N (May 18th, 2018). Towards Data Science: People Analytics with Attrition Predictions. It explains the topic as well as the logic approach using R. It helped us to analyze how to clean the data and look for correlations among variables.
* Rohan’s Four - Rohan Jain, Ali Shahid, Sehrish Saud, Julian Ramirez. IBM HR Analytics Employee Attrition & Performance. It helped us as a guideline to understand how we could approach the database using R, like how to split the data, sampling, etc.
* Baker B (February 27th, 2017). Predicting Employee Attrition. It gave us an idea of different libraries we might need to download and use to do deeper analysis.
* Sekar, A (November 30th, 2018). 5 Ways You Can Reduce Attrition Using HR Analytics. It allowed us to get a non-technical reference but support our analysis to complement and prepare conclusions and suggestions.

**Project Approach**

Like we mentioned in the proposal report, With the Kaggle dataset for IBM HR Attrition report, have clean, analyzed, and understand the dataset following these steps:

* Review Dataset

This is the first and one of the most important steps for analytical methods, we needed to understand the data to determine the best approach to achieve the objectives we defined at the beginning. In order to achieve this, we needed to:

* + Review the dataset to identify the dimensions of data available to us: like how many variables, and records, as well as understand the information.
  + Determine the key variables of interest to answer our project objectives.
* Clean Dataset

We also needed to make sure our data is clean in order to get the most accurate resources, if we had missing values, we need to replace them to avoid errors, in this case we didn’t have missing values.

Also, this dataset is quite big, so we need to understand which variables are significant for our analysis. To achieve our goals we follow the following steps:

* + Identify missing values and replace as required.
  + Drop variables that will not affect our analysis, as required.
  + Identify the dependent variables (if any)
* Exploratory Analysis

To analyze the structure of data, we need to do some exploratory data analysis. We wanted to find out the distribution of our variables to determine which ones we can use to analyze employee attrition, so we will follow these steps:

* + Perform basic transformation to analyze the dataset and dimensions
  + Visualize the variables of interest to identify trends or patterns of data
* Regression Modelling

We will analyze the correlation and predict results, for that we are planning to:

* + Develop logistic regression model with key variables to perform analysis
  + Validate the model with R-square fit/accuracy, residual analysis
* Evaluation of regression model

As importance of all the above steps, the evaluation explains how our analysis explain the employee attrition and support the decision-making process, to achieve this we will:

* + Analyze regression model values to conclude relevant correlations

With the above approach, we will be able to draw initial conclusions regarding our project questions about employee attrition.

We have been working in understand and analyze the data to determine the best approach and techniques, you can reference our Appendix to see part of our R script.

**Exploratory Data Analysis**

We’ve completed our R-script which will produce correlation and regression. The analysis is under progress. The goal is not only describing whatever is found, but also connecting the dots to show a broader picture. That is where the challenge lies.

We start with high level summary statistics to have a general understanding of what the variables’ nature. Then we plot multiple boxplot among variables with Attrition Level. Some of the plots suggest that our proposed hypothesis is correct. For example, attrition might occur more on junior professionals than senior ones. At the same time, since the junior employees have on average lower monthly income than the senior, an attrite employee is correlated with low income as well.

So far, our R-script has produced more than 8 useful plots that help us understand more about the relationship between variables. Along the analysis, we hope to identify an important relationship that we will decide to further the investigation since. However, identify leading indicator will be challenging as it requires both intuition and technical analysis.

We performed various exploration of data to go over how the data is distributed. Some interesting observations are listed below:

* The dependent variable is not evenly distributed, Attrition (No) = 1233 vs Attrition (Yes) = 237
* We compared the mean age of those leaving the company and staying the company, the
* One interesting piece from the exploration was how the attrition was among the female employees, although all female employees were distributed similar with the whole sample (No = 501, Yes = 87), the Female employees travelling frequently were behaving differently (No = 87, Yes = 30) i.e. there was higher attrition rate among this group of employees.
* We also checked other factors like employee satisfaction for both groups: attrition “Yes” and “No”, were we could check that in some variables we didn’t have major differences between one group and the others.



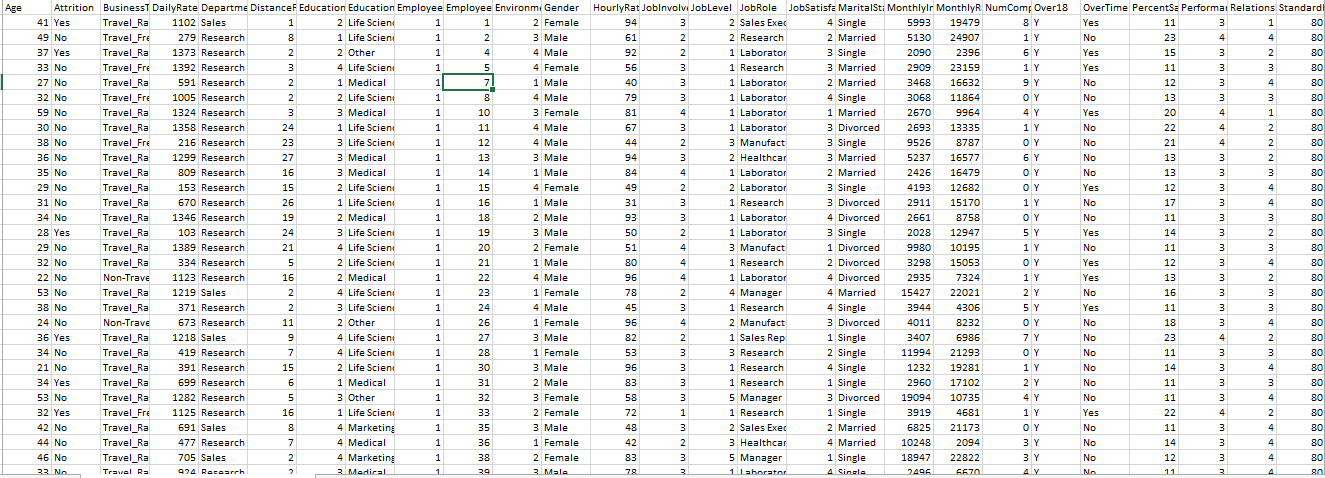
At this point, we could confidently say that age and monthly income are the important factors that impact attrition. We’ll conduct more slice and dice analysis to determine if the other variables are too.

A foreseeable risk in this exploratory data analysis step is weak result that will not be sufficient to prove our research question due to noise or unbalanced data.

**Tools & Techniques**

As we have mentioned before, our first step in all projects was clean, analyze and understand the data. Our data frame contains 1,470 records of employees and 35 variables. It contains information about personal details like age and education, as well as details about their job position, job satisfaction, travelling time, etc. Also, the data set indicates if the person is still in the company or left the company. The response in this case is the employee attrition, and the idea with this project is to understand how other variables impact in the decision of the people to stay or leave the company.

First, we were able to validate that this database does not have missing fields, so we don’t have to assume any value.



Then, we ran some information to understand the data set, like summary, length, str(), and we could start to make some assumptions about how many employees answered the questions, how many employees work for each department, how satisfied they are, gender, etc. this will help us in the future to understand how drive the employee attrition in IBM. We have 32 predictors and 1 dependent variable to begin with.

As discussed, the goal of this study is to build a model to be able to predict if a certain employee will leave the company or not. The model will be developed using the logistic regression algorithm.

* Check Assumptions

Logistic regression is still a linear regression model with the function modified to predict 0 & 1 (binary logistic regression). Hence, the assumptions of Logistic Regression are similar to the Simple/Multiple Linear Regression. We will have to make sure we satisfy all these assumptions of the Logistic Regression:

* Linearity
* Independence of errors
* Multi-collinearity

To check the linearity of the dependent variable, we started by exploring the dependent variable against the predictors and making sure it is balanced, unbalanced dataset can be fixed by up sampling or down sampling. The variance of the continuous predictors against the dependent variable would be helpful to check our independence of errors test. Also, we can develop a linear regression model among the continuous predictors to look for multi collinearity.

* Build Model:

It is recommended to start with a basic model while developing a logistic model to make sure the model (errors) converge (to global minima) which could be blocked by over complicated models. A logistic regression can be achieved in R by using the glm function from the stats package:

glm(formula, family = gaussian, data, weights, subset, na.action, start = NULL, etastart, mustart, offset, control = list(...), model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, ...)

**Challenges**

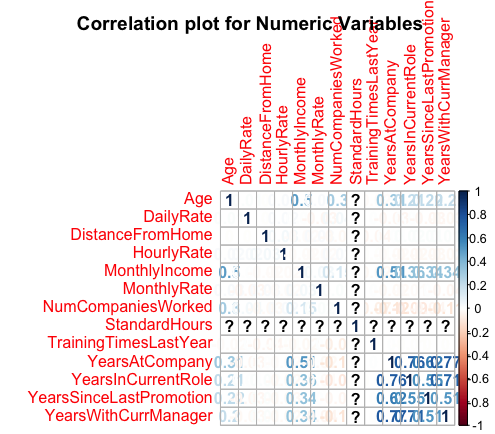
With this approach, we encountered the following challenges as we evaluated our dataset and tried to answer our research questions:

* Identifying key variables

Our dataset has 1470 observations with 35 unique variables. As we see, the sample size is small whereas the variable count is quite high. Thus, it is challenging to identify the key variables that will have any meaningful relationship with our dependent variable, i.e. - employee attrition.   
To overcome this challenge, we performed primary exploratory analysis by applying basic transformations and visualizing the data to identify trends. This is a key factor to get the best model to predict the attrition.

* Highly correlated variables

As we started analyzing the relationships of the variables, we found that quite a few variables were highly correlated. Thus, these variables show significant relationship not only with the response variable but also each other. Including these variables in our regression model might be distort results.



* Samples are imbalanced

When trying to compare data with respect to the response variable, we noted that the sample sizes for various states of the response variable are imbalanced.

**Reference**

* Kaggle (2016) IBM HR Analytics Employee Attrition & Performance. Retrieve from Kaggle (November 26th, 2018) https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
* Baker B (February 27th, 2017). Kaggle: Predicting Employee Attrition. Retrieved from Internet (December 31st, 2018): https://www.kaggle.com/brettcbaker01/predicting-employee-attrition
* Sharma, N (May 18th, 2018). Towards Data Science: People Analytics with Attrition Predictions. Retrieved from Internet (December 28th, 2018): https://towardsdatascience.com/people-analytics-with-attrition-predictions-12adcce9573f
* Rohan’s Four - Jain, R. Shahid, A. Saud, S. Ramirez, J. IBM HR Analytics Employee Attrition & Performance. Retrieved from Internet (December 30th, 2018): http://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/IBM\_Attrition\_VSS.html
* Sekar, A (November 30th, 2018). 5 Ways You Can Reduce Attrition Using HR Analytics. Retrieved from Internet (January 03rd, 2019): https://analyticstraining.com/use-hr-analytics/